

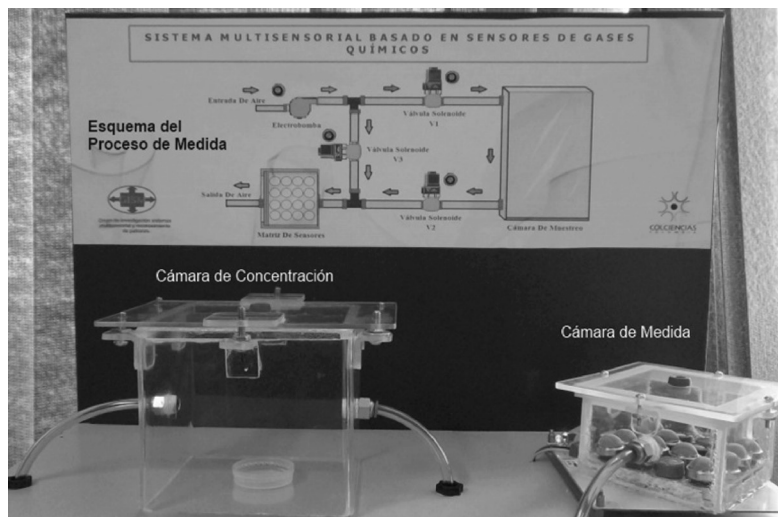
Optimización de una nariz electrónica a través de un método de selección de variables

CRISTHIAN MANUEL DURÁN ACEVEDO*

Resumen

El presente artículo aborda el proceso de optimización en la adquisición de un sistema de percepción artificial (nariz electrónica), el cual está compuesto de una matriz que cuenta con 16 sensores de gases químicos y métodos de procesamiento de señales para la clasificación de compuestos químicos. Se utilizaron en total tres compuestos orgánicos volátiles (COV), de tipo hidrocarburos aromáticos (benceno, tolueno y xileno), a diferentes niveles de concentración, entre los 500, 1000 y 1500 ppm (partes por millón). El objetivo del estudio fue mejorar el funcionamiento de una nariz electrónica, a través de la implementación de métodos de reconocimiento de patrones (MLP, PNN, SVM), y la técnica de selección de variables 'Simulated annealing (SA)' y PNN (Probabilistic Neural Network), obteniendo una reducción importante de la matriz de datos y llegando a un porcentaje de acierto de hasta el 100% en la clasificación.

Palabras claves: nariz electrónica, compuestos químicos, selección de variables, procesamiento, sensores de gases.



(*) Facultad de Ingeniería y Arquitectura, Universidad de Pamplona, Bucaramanga - Colombia. Grupo de Investigación en Sistemas Multisensoriales y Reconocimiento de Patrones.

Fecha de recepción: 07/03/2013 • Fecha de aceptación: 30/04/2013.

Abstract

The present article consists in optimization of acquisition of an artificial perception system (electronic nose), which is composed of an array of 16 gas sensors chemicals and signal processing methods for classifying chemical compounds. A total of three volatile organic compounds (VOC) were used such as aromatic hydrocarbons (benzene, toluene and xylene) at different concentration levels of 500, 1000 and 1500 ppm (parts per million). The goal of study was to improve the performance of an Electronic Nose through the implementation of pattern recognition methods (MLP, PNN, SVM) and the variable selection technique 'Simulated annealing (SA)' and PNN (Probabilistic Neural Network), obtaining a reduction in the matrix data and reaching a success rate of 100% in classification.

Keywords: electronic nose, chemical compounds, variable selection, processing, gas sensors.

1. Introducción

En los sistemas de análisis de aromas como las 'narices electrónicas', por lo general se detectan compuestos a partir de una matriz de sensores de gases químicos con sensibilidades parcialmente sobrepuestas. Por lo tanto, con relación al concepto de la nariz electrónica, es necesario tener en cuenta que un número significativo de sensores puede ser irrelevante, puesto que, generalmente, algunos de los sensores presentan con el tiempo desgaste o derivas generadas por cambios en las variables físicas o eléctricas del dispositivo; por ejemplo: cambios de temperatura, humedad o variación eléctrica. En forma general y para diferentes aplicaciones, se puede considerar que otros sensores pueden presentar buen comportamiento en cada una de las respuestas.

En la actualidad diferentes estrategias se han reportado para la reducción de la dimensionalidad en el conjunto de datos. Esta reducción consiste básicamente en la selección de cualquier variable disponible en dicho conjunto o en la generación de nuevas variables llamadas factores (por ejemplo, mediante la realización de análisis de componentes o de un análisis discriminante

lineal, etc) (Alizadeh, 2010). Cuando el sensor presenta variaciones importantes, es primordial implementar métodos estadísticos (por ejemplo: Análisis de Componentes Principales-ACP), como un paso previo en la reducción de dimensionalidad en el conjunto de datos.

Para el caso del método de ACP, se realiza un análisis a través de un plano bidimensional, el cual conlleva a seleccionar los sensores mediante un criterio de análisis de scores y loadings, que consiste en eliminar los sensores que estén más cerca del valor cero del plano y en seleccionar los sensores que se agrupen en un punto específico del plano para realizar también dicha eliminación (Gardner & Barlett, 1999).

Otras estrategias han sido desarrolladas para selección de variables, las cuales conducen a la construcción de modelos de clasificación parsimoniosa basada en redes neuronales tipo 'Fuzzy ARTMAP' o redes neuronales probabilísticas. Este proceso consiste en seleccionar variables mediante dos etapas; la primera a través de una selección basada en un criterio de varianza, seguido por un proceso de recocido simulado (Llobet, Brezmes, Gualdrón, Vilanova & Correig, 2004; Llobet et al., 2007).

En otra aplicación un algoritmo de búsqueda estocástica, por medio de Algoritmos Genéticos (AG), utiliza ciertas funciones características como un primer paso para optimizar problemas de la selección del sensor con mayor grado de información (Phaisangittisagul, Nagle & Areekul, 2010).

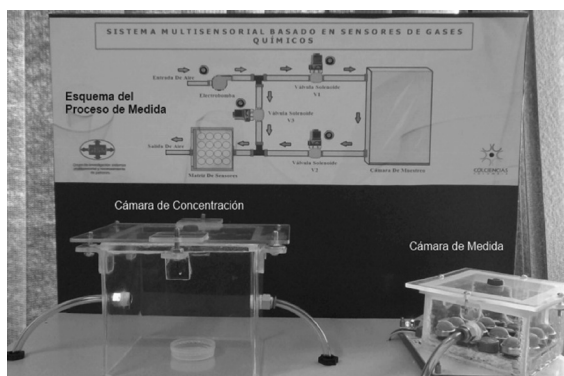
El módulo desarrollado en este proyecto cuenta con una matriz de 16 sensores, que permite obtener un gran conjunto de datos por cada medida. Además presenta la utilidad de realizar diferentes pruebas de laboratorio en el sector agroalimentario e industrial (ejemplo: control de calidad de los alimentos) y puede ser implementado en diferentes sectores, como en el caso de la salud. Al obtener un elevado número de sensores, el sistema puede generar un problema en el que se obtiene una matriz con gran cantidad de información, donde existen variables que aportan información útil y otras que solo aportan ruido. Por ese motivo, es importante escoger cuidadosamente las variables (es decir, los sensores en este caso), a través de una técnica de selección y de reconocimiento de patrones que puedan ser utilizados para obtener un buen funcionamiento del sistema.

En este artículo se presenta la optimización de una nariz electrónica, mediante la implementación de una técnica de selección de variables (es decir, SA) y la utilización de diferentes métodos de reconocimiento de patrones, como las redes neuronales MLP, PNN y el clasificador SVM (Gualdrón, Durán, Isaza, Carvajal & Uribe, 2011).

Estos métodos fueron comparados mediante la aplicación de la técnica de selección de variables ‘recocido simulado’ a cada uno de estos métodos, obteniendo una importante reducción de dimensionalidad en la matriz de datos y llegando a un porcentaje de acierto en la clasificación de las medidas de hasta el 100%, en algunos casos.

2. Materiales y métodos

Figura 1. Fotografía del equipo multisensor



Fuente: elaboración propia.

El esquema del proceso de medida lo componen tres etapas principales. Una etapa de concentración (es decir, de espacio de cabeza estático) con flujo constante, una matriz de 16 sensores químicos tipo TGS —ubicada en la cámara de medida— y un sistema informático. En la Figura 1 se muestra el esquema general del equipo multisensor, en el cual se pueden distinguir los diferentes elementos que conforman el equipo. El procedimiento de medida en el método consta de tres fases bien diferenciadas: la fase de concentración, la fase de medida y la fase de reposo. Los tiempos de medida fueron los siguientes:

A. Tiempo de Concentración (15 minutos): es el momento en que se produce un aumento en la concentración de las muestras, debido al acondicionamiento en la

cámara de concentración. Los volátiles no recorren la cámara de medida donde están ubicados los sensores. En este caso se genera una línea base constante en la respuesta de los sensores.

B. Tiempo de Medida (10 minutos): es el tiempo en el que los volátiles llegan a la cámara de medida, con el fin de que los sensores respondan adecuadamente a la muestra.

C. Tiempo de Reposo (20 minutos): los sensores se llegan a estabilizar por un tiempo dado, llegando a recuperar la línea base, que debe ser aproximada a la fase de concentración.

2.1 Cámara de concentración

En la cámara de concentración una cubierta superior de acrílico fue equipada de un septum, con el objetivo de inyectar manualmente los compuestos químicos por medio de una jeringa milimétrica.

Las dimensiones de la cámara de concentración fueron las siguientes: 12 cm. de largo x 12 cm. de ancho x 12 cm. de alto, para un volumen total de 1728 cm³. Las dimensiones de la cámara de medida fueron de 8.7 cm. de largo x 8.7 cm. de ancho x 5.3 cm. de alto, para un volumen total de 401.157 cm³, la cual también fue desarrollada en acrílico transparente. El sistema de muestreo, en general, fue acondicionado empleando la cámara de concentración, la cámara de medida, cuatro electroválvulas solenoides y una bomba de aire de 5 bares de presión para transportar los volátiles.

2.2 Sensores de gases

Los 16 sensores de gases utilizados para el desarrollo del equipo fueron de tipo semiconductor de la casa Figaro Inc., concretamente de óxido de estaño (Figaro, 2012).

En la Tabla 1 se describen cada uno de los sensores de gases utilizados en la cámara de medida.

Tabla 1. Sensores utilizados en el método

Sensor de gas Serie-8	Aplicación	Tipo de gas	Número
TGS 826	Detección de olores	Amoníaco y aminas	2
TGS 831	Detección de gas halocarburos (gases refrigerantes)	R-22 Monoclorodifluorometano	2
TGS 821	Detección de gas combustible	Hidrógeno	2
TGS 842	Detección de gas combustible	Metano y gas natural	2
TGS 880	Control de cocción	Humo de los alimentos (alcohol, olor)	2
TGS 825	Detección de gas tóxico	Sulfuro de hidrógeno	1
TGS 813	Detección de gas combustible	Hidrocarburos en general	1
TGS 800	Control de calidad del aire	Contaminantes del aire en general	1
TGHS 822	Detección de vapores de disolventes	Alcohol y disolventes orgánicos	1
TGS 832	Detección de gas halocarburos (gases refrigerantes)	R-134 1, 1, 1, 2-Tetrafluoroethane	1
TGS 830	Detección de gas halocarburos (gases refrigerantes)	R-22 Monoclorodifluorometano	1

Fuente: elaboración propia.

Para adquirir y almacenar las señales de los sensores al ordenador, se utilizó una tarjeta de alto rendimiento del fabricante National Instruments, la DAQ, NIUSB-6210, que es un módulo multifunción de la Serie M, optimizado para una precisión superior a velocidades de muestreo muy altas.

2. 3 Métodos de selección de variables y procesamiento de datos

A continuación se describe brevemente el método de selección de variables y las técnicas de reconocimiento de patrones utilizadas en la clasificación de compuestos y reducción de dimensionalidad del conjunto de datos.

2. 3. 1 Simulated annealing o recocido simulado

Simulated annealing es una técnica estocástica de optimización que permite hallar soluciones cercanas al óptimo global en problemas de optimización complejos (por ejemplo. con elevado número de variables). El algoritmo parte de la utilización de un conjunto de variables (generalmente todas las variables disponibles), X_0 . Una vez calculado el fitness (error de predicción obtenido con la red clasificadora PNN) asociado a X_0 , se determina un conjunto X_1 , que resulta después de haber eliminado n variables (escogidas aleatoriamente dentro de X_0). El fitness de

X_1 es obtenido y se calcula la diferencia entre el modelo original y el nuevo.

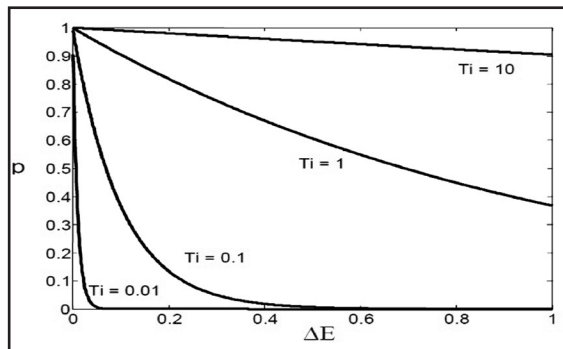
$$\Delta E = \text{fitness}(\text{nuevo}) - \text{fitness}(\text{viejo}) \quad (1)$$

Se puede afirmar que ΔE es negativo si el nuevo modelo es superior al original, es decir, cuando el resultado del cálculo del error de predicción es menor, por lo que la clasificación ha mejorado eliminando dicha variable. Ahora bien, si el modelo resultante al utilizar las variables de X_1 es peor al original (ΔE positivo), no significa directamente que dicha combinación deba ser rechazada. Se pasaría a una segunda fase en la que se define la probabilidad:

$$p_i = \exp(-\Delta E/T_i) \quad (2)$$

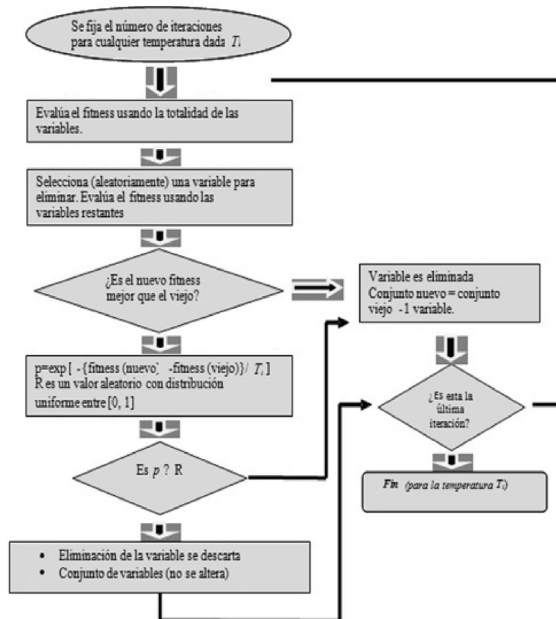
Donde T_i es la temperatura de trabajo (se escoge un valor inicial para T_i). Si $p_i > R$ (R es un valor aleatorio con distribución uniforme entre 0, 1) la nueva solución es retenida y el algoritmo prosigue eliminando variables a partir de X_1 . En el caso contrario, el algoritmo prosigue desde X_0 . La Figura 2 muestra a p como una función de los cambios de fitness (valores positivos de ΔE) a diferentes temperaturas annealing T_i . Cuando la temperatura de annealing se reduce, la posibilidad para aceptar una solución peor decrece significativamente. El proceso se ejecuta en un número determinado de iteraciones ($\#_{\text{iteraciones}} = N_{\text{variables totales}} - 1$) para la temperatura T_i . Finalmente todo lo anterior se repite para $T_{i+1} = \alpha T_i$ ($\alpha < 1$). En otras palabras, el proceso de selección se desarrolla a una temperatura más baja (el proceso de aceptación de cambios es menos exigente). El número de temperaturas a computar también debe ser definido *a priori*, teniendo en cuenta que un número bajo puede dar como resultado soluciones no muy buenas y un número alto incrementaría drásticamente el tiempo de computación del proceso (un valor promedio de ejecución puede ser de 50 temperaturas). En la Figura 3 se puede observar el diagrama de flujo del algoritmo simulated annealing. Este algoritmo se utilizó para la red neuronal PNN, principalmente, porque arrojó resultados satisfactorios.

Figura 2. Probabilidad p de aceptación a soluciones peores como una función del cambio de *fitness* a diferentes temperaturas *annealing*, T_i



Fuente: elaboración propia.

Figura 3. Diagrama de flujo del algoritmo *simulated annealing*



Fuente: elaboración propia.

2. 3. 2 Métodos de reconocimiento de patrones

A continuación se relacionan los métodos de reconocimiento de patrones utilizados para el procesamiento de información, los cuales fueron aplicados a la matriz de datos para la clasificación de las medidas:

- MLP (Multilayer Perceptron).
- PNN (Probabilistic Neural network).
- SVM (Support Vector Machine).

- SVM Multiclase.
- LS-SVM (Least squares -Support Vector Machine).

La Tabla 2 resume las principales características de los algoritmos aplicados.

Tabla 2. Principales características de los algoritmos aplicados al sistema

Técnica	Algoritmo	Aprendizaje	Par	Aplicación
MLP	Neuronal	Supervisado	No	-Clasifica - Cuantifica
PNN	Neuronal	Supervisado	No	Clasificar
SVM	Obtiene una frontera de decisión	Supervisado	No	Clasificar y análisis de regresión
Multi-clase SVM	Obtiene una frontera de decisión. Se construyen N modelos SVM y $N(N-1)/2$ clasificadores	Supervisado	No	Clasificar análisis de regresión
Least squares - Support Vector Machine LS SVM	Obtiene una frontera de decisión con mínimos cuadrados	Supervisado	No	Clasificar análisis de regresión

Fuente: elaboración propia.

En la Tabla 3 se describen los métodos usados para extraer los parámetros estáticos más relevantes del conjunto de datos. Estos parámetros fueron aplicados a valores de conductancia ($1/R_s$). La conductancia fue obtenida a partir de los valores de resistencia (R_s) de la señal de salida en cada uno de los sensores.

Tabla 3. Métodos de extracción de parámetros

Parámetro	Expresión	Comentario
Máximo incremento de la conductancia	$G_{max} - G_{min}$	Minimización del efecto de la temperatura
Incremento de conductancia	$G_f - G_i$	
Máximo incremento de la conductancia normalizada o máxima diferencia fraccional	$(G_{max} - G_{min}) / G_{max}$	
Incremento de conductancia normalizada o incremento de la diferencia fraccional	$(G_f - G_i) / G_f$	

Fuente: elaboración propia.

En la Tabla 4 se describen las diferentes técnicas utilizadas para el pre-procesado de datos, las cuales realizan tareas de normalización y auto-escalado.

Tabla 4. Métodos de pre-procesado de datos

Normalización	Descripción
Auto-escalado (<i>autoscaling</i>)	El objetivo de esta normalización es la de dar igualdad de escala a cada una de las variables o parámetros que describen cada medida.
Centrado (<i>mean centring</i>)	En esta normalización a cada una de las columnas (variables) se le resta su valor medio. Es decir, toda medida es descrita por variables de media nula.
Normalizado o normalizado por matriz	Divide toda la matriz por el elemento máximo de la misma. En este caso solo un elemento valdrá la unidad.

Fuente: elaboración propia.

2. 4 Muestras

Las muestras fueron preparadas con un volumen de 0.1 ml. sobre un recipiente de plástico previamente esterilizado, para garantizar que cada prueba se hiciera de forma igual y homogénea. La extracción de la muestra del recipiente contenedor se realizó mediante una jeringa milimétrica etiquetada para cada compuesto. En total se hicieron 45 medidas con densidades aproximadas del 0.88 g/ml. y un promedio de 870.000 ppm de concentración.

Relación de medidas:

- 15 medidas de benceno.
- 15 medidas de tolueno.
- 15 medidas de xileno.

3. Resultados

A continuación se describen los resultados obtenidos a través de la clasificación de compuestos aromáticos (benceno, tolueno y xileno) a diferentes niveles de concentración de 500, 1000 y 1500 ppm, mediante el algoritmo de selección de variables (SV) ‘Simulated Annealing (SA)’, acoplado a la red neuronal PNN. En todos los casos se realizó la clasificación de medidas con el método de normalización ‘centrado de datos’.

En el conjunto de datos se aplicó el SVM (uno contra todos) en todas las clases, donde se construyeron N modelos. El SVM fue entrenado con todas las muestras de entrenamiento de la clase con etiquetas (*labels*) de valores positivos y todas las otras muestras de entrenamiento con *labels* negativos. Una nueva medida pertenece a la clase que tiene los valores altos de la función indicador.

Se debe aclarar que el método de selección de variables fue SA-PNN, el cual fue aplicado a la matriz de 16 sensores de gases, generando mejores resultados con información relevante del conjunto de datos, para luego ser aplicados a cada uno de los métodos de clasificación.

3. 1 Clasificación de todos los compuestos con el método de selección SA-PNN

La Tabla 5 ilustra los porcentajes de aciertos en la clasificación de las muestras de benceno, tolueno y xileno, usando redes neuronales. SVM multiclases y SVM mínimos cuadrados, con el procedimiento uno contra todos con selección de variables.

Tabla 5. Clasificación de todos los compuestos usando redes neuronales. SVM multiclases y SVM

Parámetro	Método de selección	Número de variables	Método de clasificación	% con SV
Gmax-Gmin 45x16	SA-PNN	2,3,5,8,11,12	MLP	100
			PNN	100
			SVM Multiclases	67
			LS SVM Multiclases	100

Fuente: elaboración propia.

Aplicando el parámetro estático a la matriz de 16 sensores de gases y con el método de selección de variables, se seleccionaron en total 7 sensores. Con la mayoría de los métodos de reconocimiento de patrones se obtuvo 100% de acierto en la clasificación de las medidas.

3. 2 Clasificación del benceno con el método de selección SA-PNN

La Tabla 6 ilustra los porcentajes de aciertos en la clasificación de las muestras de benceno a 1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM mínimos cuadrados.

Tabla 6. Clasificación de las muestras de benceno con SVM multiclases y SVM mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro	Método de selección	Número de variables	Método de clasificación	% con SV
Gmax-Gmin 45x16	SA-PNN	2,3,4,10, 11,13,16	MLP	100
			PNN	80
			SVM Multiclases	67
			LS SVM Multiclases	100

Fuente: elaboración propia.

Al igual que el resultado anterior, se logró disminuir la matriz a solo 7 sensores de gases, disminuyendo un 20% de acierto en la clasificación, mediante el uso de las redes PNN.

3. 3 Clasificación del tolueno con el método de selección SA-PNN

La Tabla 7 ilustra los porcentajes de aciertos en la clasificación de las muestras de tolueno a 1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM mínimos cuadrados.

Tabla 7. Clasificación de las muestras de tolueno con SVM multiclases y SVM mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro	Método de selección	Número de variables	Método de clasificación	% con SV
Gmax-Gmin 45x16	SA-PNN	1,2,3,4,9, 12	MLP	100
			PNN	93
			SVM Multiclases	33
			LS SVM Multiclases	97

Fuente: elaboración propia.

Para realizar la clasificación de las concentraciones de las muestras de tolueno se utilizó en este caso una matriz de 45x6 sensores, obtenido de esta manera un 100% de clasificación a través de la red MLP.

3. 4 Clasificación del xileno con el método de selección SA-PNN

La Tabla 8 ilustra los porcentajes de aciertos en la clasificación de las muestras de xileno a

1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM mínimos cuadrados.

Tabla 8. Clasificación de las muestras de xileno con SVM multiclases y SVM mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro	Método de selección	Número de variables	Método de clasificación	% con SV
Gmax-Gmin 45x16	SA-PNN	1,2,4,1,3, 15,16	MLP	100
			PNN	93
			SVM Multiclases	33
			LS SVM Multiclases	97

Fuente: elaboración propia.

Mediante la última prueba realizada a las concentraciones de xileno, se seleccionaron en forma similar un número total de 7 sensores de gases, llegando a obtener el mismo porcentaje de clasificación con las muestras de tolueno, pero con la diferencia de un sensor a favor.

4. Conclusiones

En este estudio realizado a través de un prototipo de nariz electrónica, se concluye que este tipo instrumentos de medida puede ser optimizado a través de métodos computacionales, mediante una estrategia basada en técnicas específicas de selección de variables y métodos de reconocimiento de patrones. Se logró reducir la dimensionalidad de una cámara de medida compuesta por una matriz de 16 sensores de gases químicos a través de los métodos de procesamiento, llegando a alcanzar un valor del 50%, aproximadamente.

El método de selección de variables SA-PNN utilizado en el conjunto de medidas adquiridas previamente por la nariz electrónica, permitió eliminar un número determinado de variables (es decir, sensores) redundantes o ruidosas, mejorando así, en la mayoría de los casos, el porcentaje de clasificación del número de muestras a partir de las respuestas de cada uno de los clasificadores. Por ejemplo, para el caso de la clasificación de todos los compuestos (xileno, benceno y tolueno) —correspondiente a 45 medidas—, se obtuvo una clasificación del 100% de acierto al utilizar las redes neuronales (como la MLP y PNN), elevando

hasta un 10% más de clasificación y eliminando una cantidad considerable de variables.

En la clasificación de las 15 medidas de cada uno de los compuestos químicos, se llegó a obtener también un 100% de acierto en la clasificación de las medidas, mejorando el funcionamiento de la nariz electrónica, en cuanto a la reducción de la matriz de sensores; así mismo, optimizando de forma significativa el sistema sensorial y el hardware de adquisición de datos.

Es importante aclarar que al encontrar y utilizar el mejor método de clasificación, es posible reducir la dimensión de la matriz de sensores de gases, generando de esta forma una importante optimización con respecto a los materiales utilizados para el desarrollo de este tipo de sistemas sensoriales.

Agradecimientos

El grupo de Investigación en Sistemas Multi-sensoriales y Reconocimiento de Patrones de la Universidad de Pamplona, agradece al ingeniero electrónico Luis Daniel Rojas, por su colaboración y excelente desempeño en este trabajo.

Los resultados obtenidos en este artículo fueron llevados a cabo gracias al proyecto cofinanciado por Colciencias y la Universidad de Pamplona. ●

Referencias

Alizadeh, T. (2010). Chemiresistor sensors array optimization by using the method of coupled statistical techniques and its application as an electronic nose for some organic vapors recognition. *Sensors and Actuators B: Chemical*, 143, (2), 740-749.

Figaro. (2012, enero). *Information for TGS Sensors*. Recuperado de <http://www.figarosensor.com/products/general.pdf>.

Gardner, J. & Barlett, P. (1999). *Electronic Noses: Principles and Applications*. Oxford University, New York. pp. 154-155.

Gualdrón, O., Durán, C., Isaza, C., Carvajal, A. & Uribe, C. (2011). Sistema de olfato electrónico de bajo costo para la detección de diferentes compuestos químicos contaminantes. *Revista Colombiana de Tecnologías de Avanzada*, 1, (17), 121-126.

Llobet, E., Gualdrón, O., Vinaixa, M., El-Barbri, N., Brezmes, J., Vilanova, X., Bouchikhi, B., Gómez, R., Carrasco, J. & Correig, X. (2007). Efficient feature selection for mass-spectrometry based electronic nose applications. *Chemometrics and Intelligent Laboratory Systems*, 85, 253-261.

Llobet, E., Brezmes, J., Gualdrón, O., Vilanova, X. & Correig, X. (2004). Building parsimonious fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm; application to multisensors systems for gas analysis. *Sensors and Actuators B*, 99, 267-272.

Phaisangittisagul, E., Nagle, H.T. & Areekul, V. (2010). Intelligent method for sensor subset selection for machine olfaction. *Sensors and Actuators B: Chemical*, 145, (1), 507-515.